DOCUMENT RESUME

ED 311 075 TM 013 912

AUTHOR Busch, John Christian; Jaeger, Richard M.
TITLE An Exploratory Assessment of Content Validation Processes Used with the National Teacher Examinations.
SPONS AGENCY North Carolina Univ., Greensboro. Center for Educational Research and Evaluation.
PUB DATE Mar 89
NOTE 43p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Content Validity; Data Collection; *Evaluators; Formative Evaluation; Higher Education; National Programs; *Professors; *Public School Teachers; Relevance (Education); Science Tests; Self Evaluation (Individuals); Standardized Tests; *Teacher Certification; Teacher Education; Testing Programs; Test Validity
IDENTIFIERS *Experts; *National Teacher Examinations

ABSTRACT

The role of expert judges in establishing the content validity of the National Teacher Examinations (NTE) was examined in a detailed study. The NTE are used in 22 states for screening applicants to teacher education programs and/or for screening candidates for initial teacher certification. Such extensive use of the NTE has stimulated 35 recent validation and standard-setting studies of the tests. Most of the studies have sought the judgments of experts concerning the relevancy of the tests to success in a teacher education program or to success as a beginning teacher. This study assessed characteristics and selection of judges, data collection procedures for judgment panels, validity of judges' instruction concerning student opportunity-to-learn assessments, judges' attitudes toward the uses of the NTE, and their personal confidence in their judgments. This paper reports results only for the seven multiple-choice subtests that compose the communication skills and general knowledge tests of the NTE. In this study, 82 college/university teachers and 37 public school teachers assessed the NTE communication skills subtest, and 87 college/university teachers and 39 public school teachers assessed the NTE general knowledge subtest. Results indicate that: (1) 21% of the judges did not agree that applicants should be screened at all; (2) 72% of the judges felt that tests should be validated prior to use; (3) assessments of the relevancy of examinees' opportunity to learn the content of the NTE were generally positive; (4) the science subtest was rated the lowest with respect to relevancy; and (5) judges were generally confident about their judgments. Seven data tables and three bar graphs are provided. (TJH)

ED311075

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JOHN CHRISTIAN BUSCH
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Exploratory Assessment of Content Validation Processes Used With The National Teacher Examinations *

John Christian Busch and Richard M. Jaeger
University of North Carolina at Greensboro

Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, March 1989

* Supported in part by the Center for Educational Studies and Development, University of North Carolina at Greensboro

TM013912

The National Teacher Examinations (NTE) are used in 22 states for screening applicants to teacher education programs and/or for screening candidates for initial teacher certification (Rudner, 1987). Such extensive use of the NTE has stimulated thirty-five recent validation and standard-setting studies of the tests (Cross, 1985).

Although some researchers have investigated the criterion-related validity (e.g., Brown & Rankin, 1986) and construct validity (e.g., Rosenfield, Thornton & Skurnick, 1986; Rosner & Howey, 1982) of the NTE, most recent studies have sought the judgments of "experts" on the relevance of the tests to success in a teacher education program or to success as a beginning teacher, depending on the proposed use of the NTE. Such studies have been termed "content validity assessments". In addition, some studies have sought experts' judgments on the candidates' opportunities to learn the material assessed by the NTE (sometimes termed "instructional validity").

Messick (1989), paraphrasing the 1954 and 1966 *Test Standards*, states that "Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are drawn" (p. 16), and observes that content validity is "based on professional judgments about the relevance of the test content to the content of a particular behavioral domain of interest and about the representativeness with which the item or test content covers that domain" (p. 17). Although the issue of content representativeness has been examined in some validity studies of the NTE, recent writers (e.g., Carlson, 1988) have suggested that it is important to establish the *necessity* of the knowledge or skill measured by the test to successful performance in a relevant domain. This focus is consistent with the prescription in the

Uniform Guidelines on Employee Selection Procedures (1978) that evidence be provided in support of the "job relatedness" of the test.

Different subtests of the NTE have been used in making initial teacher certification decisions, and in screening applicants to teacher education programs. The domain of interest is different in each application. In the first application, the domain consists of knowledge and skills that are essential to successful performance as a beginning teacher; in the second application, the domain consists of knowledge and skills essential to successful performance as a student in a teacher education program.

The Opportunity to Learn.

Messick (1989) notes that "the judged relevance to the curriculum...has been deemed insufficient for claiming that a test is relevant for educational certification because the curriculum might not have been fully implemented for the students" (p. 67). He refers to the need to provide evidence of instructional validity, which he describes in the following terms, "perverse", "duplicative", "vague", and "refractory to appraisal". Although Messick takes issue with the use of the term "validity", he acknowledges the need to assess the relationship between test scores and examinees' instructional histories.

Jaeger (1989) analyzes the psychometric and legal history of the "opportunity to learn" concept. He concludes, as does Messick (1989), that content validity is not part of the "validity" construct. He prefers using the term "opportunity to learn" as an umbrella expression. Opportunity to learn is an issue of "fairness"; individuals should not be tested on knowledge which they have not had the opportunity to obtain, an issue successfully argued in the *Debra P. v Turlington* case. Typically, there have been two aspects of this issue, termed by McClung (1979) "curricular validity" and "instructional validity." Curricular validity concerns the congruence of test

content and curriculum content. Instructional validity concerns the congruence of test content and the skills or knowledge embodied in instruction. Jaeger (1989) prefers the terms "curricular coverage" and "instructional preparation". Whether evidence of content congruence should be required, and in what form, is a matter of some debate. For example, Cross (1988) expressed doubts concerning judges' abilities to make informed judgments about examinees' opportunities to learn skills and knowledge assessed by the less well-structured tests of the NTE (e.g., the Professional Knowledge test in contrast to the Mathematics subtest). Poggio, Miller and Glasnap (1987) found retrospective judgments of the instructional validity of the NTE to be adequate. Millman and Greene (1989) conclude that the level of evidence regarding opportunity to learn that will be required by the courts is unclear; that is, will the test user be required to demonstrate that examinees had sufficient opportunity to learn, or will the requirement be less restrictive; e.g., a demonstration that lack of opportunity was not capricious? In a study for the state of North Carolina, Jaeger and Busch (1984) collected three types of evidence bearing on examinees' opportunity to learn the content of the NTE. This evidence is described in a later section of this paper.

Expert Judges

The validation processes described above are grounded in evidence produced by expert judges. However the adequacy of evidence based on expert judgment and what qualifies one to be an "expert" are matters of discussion. In the judgment of Messick (1989), consensus judgments by experts are important, but "such judgment can be fallible and needs to be buttressed by construct related evidence as to what the items measure" (p. 39). Cross (1988) concluded, with reference to his study of the NTE for the

state of Virginia, that he had greater regard for the relevance judgments provided by public school judges than he did for those provided by college and university faculty. He felt that classroom teachers would have a better understanding of the actual practice of teaching against which to evaluate the relevance of the NTE. In a North Carolina study of the use of the NTE to screen candidates for teacher education programs, Jaeger and Busch (1984) concluded that both college and university faculty and public school teachers could reasonably serve as "experts". The present study provides an opportunity to examine differences between the validity recommend-ations furnished by these two types of judges.

Although relationships between characteristics of expert judges and their standard-setting recommendations have been examined in several studies (Cross, et al.,1985; Busch 1988; Busch & Jaeger, 1988), parallel studies of relationships between judges' characteristics and their recommendations concerning the content validity or instructional validity of the NTE have not been reported. Cross (1985) found that public school teachers and college faculty differed significantly in their mean recommended test standards. Jaeger (1982) found substantial differences in median test standards recommended by public school teachers, public school principals, and registered voters. Busch (1988) reported that public school teachers and college faculty differed significantly in the variability of the standards they recommended for the NTE. This study investigated the generalizability of these findings to test validity judgments.

Objectives of the Study

This study had four objectives: (1) to examine and report expert judges' confidence in the quality of their recommendations and their assessments of the adequacy of procedures used in a typical NTE validation study, (2) to

examine and report expert judges' attitudes toward standardized testing (in particular the NTE), and their acceptance of the need to validate such tests, (3) to investigate relationships among characteristics of judges that can be controlled by investigators either during the selection of judges (e.g., whether selected experts are public school teachers or college faculty members), or prior to analyses of judgment data (e.g., experts' reported confidence in the quality of their recommendations and judged adequacy of validation procedures), and their judgments of the content validity and instructional validity of the NTE, and (4) to investigate relationships among experts' attitudes toward testing and acceptance of the need to validate tests and their judgments of the content validity and instructional validity of the NTE.

Methods

Characteristics and Selection of Judges

The design used for sampling judges is described in detail in Busch, Jaeger, Pratto & Sullivan (1983) and in Jaeger and Busch (1984). It required that two types of judges be selected. Selection was to be completed using a probability sampling procedure that ensured unbiased representation of (a) statewide populations of public school teachers who were relatively recent graduates of approved teacher education programs and (b) college and university faculty who contributed to the education of beginning teachers and who were employed in either discipline-based departments or education departments or schools.

The design proposed that data be collected at three sites that were convenient to colleges, universities and school systems in the mountain, piedmont and coastal regions of North Carolina. The data collection sites

were selected so as to secure an effectively representative statewide sample of public school teachers who satisfied the criteria described above, without attempting the impractical task of sampling teachers from every school system in North Carolina.

The county that best represented each of the three regions of North Carolina on six demographic variables was selected from each region. Teachers were sampled from all school systems that were within a 35 mile radius of the selected county. Using this criterion, 48 school systems were identified as sources of public school participants. The State Department of Public Instruction provided sampling frames for all teachers who met the following five criteria: Full-time teacher, graduate of one of 44 approved teacher education programs in North Carolina; currently employed in a sampled school system; a holder of a valid teaching certificate as of 1983; initially certified no more than five years and no less than two years prior to 1983. Teachers were sampled using a stratified cluster sampling procedure with region as the stratification variable, and proportional representation by region.

Detailed descriptions of the development of sampling frames for college and university faculty in discipline departments who taught undergraduate general education courses in subjects assessed by the General Knowledge and Communication Skills Tests of the NTE, and faculty in departments, schools and colleges of education, are provided in Busch et al. (1983) and Jaeger and Busch (1984). The sampling design was intended to meet a number of objectives: First, to ensure representativeness, a probability sampling procedure that incorporated random selection of participants at the final stage of sampling was to be used. Second, institutions in each region were to be represented in proportion to the numbers of graduates from their teacher

education programs in recent years. Third, institutions that traditionally served minority students were to be sampled at a higher rate than were institutions that traditionally served non-minority students. The numbers of judges of each type who were to serve on the panels assembled to review the seven multiple-choice sub-tests and the Writing Essay subtest of the NTE were determined using statistical criteria that guaranteed estimation of population proportions within +_ 0.05 with 95 percent confidence. College and university faculty reviewed subtests that were appropriate to the area of their teaching responsibilities. (For example, discipline based faculty from biology, physics, and chemistry departments and teacher education faculty who taught science education courses were selected to review the NTE Science subtest). Seventy-six public school teachers and 169 college and university faculty served on seven judgment panels, excluding a panel that reviewed the Writing Essay subtest.

The procedures used to collect validity evidence for the Writing Essay subtest was different from the procedures used with the multiple-choice tests. This paper reports results only for the seven multiple-choice subtests that compose the Communication Skills (Reading, Writing, and Listening) and General Knowledge (Mathematics, Science, Social Studies, and Literature/Fine Arts) tests of the NTE. The numbers of public school teachers and college and university faculty who were members of the judgment panel that examined each NTE subtest is reported in Table 1.

Insert Table 1 about here

Data Collection Procedures

Each judgment panel was instructed by a knowledgeable group leader who was supported by an assistant. Group leaders had professional training and responsibilities in the subject area of the subtest reviewed by their respective panels and were themselves trained by the principal investigators during a five-hour training session. All data-collection procedures were standardized and were conducted using detailed scripts, handout materials, and transparencies. The validation activities were alternated with exercises used to establish test standards; data collection lasted a full day. Only validation activities will be described in this paper.

The first judgment session began with a review of background information on the North Carolina Quality Assurance Program and the National Teacher Examinations. Judges then "took" the subtest they were going to review. The validation activity consisted of two major parts: Assessment of the relevance of the NTE subtests and assessment of examinees' opportunity to learn the skills and knowledge measured by the subtests. Both types of judges were considered competent to make judgments regarding the relevance of the knowledge and skills assessed by the subtests to success in approved teacher education programs. Only the college and university faculty provided judgments regarding examinees' opportunity to learn the content of the subtests. College and university faculty were presumed to be familiar with the curricula and educational experiences of students in North Carolina's approved teacher education programs; public school teachers were presumed to be less familiar with these topics, and therefore not qualified to make judgments on examinees' opportunity to learn the content of the NTE subtests.

Collection of Relevance Judgments. Each judge was provided a copy of the NTE subtest (s)he was to review and an Item Relevance/Instructional Validity Response Form. Judges were asked to work independently and to make a single judgment for each item. They were to determine the skill or knowledge that was measured by the item, and then to judge how important that skill or knowledge was to success in the North Carolina teacher education program with which they were most familiar. Judges rated the importance (relevance) of the skill or knowledge on the following four-point scale:

> "Crucial" - of crucial importance to success in a teacher education program. These are skills or knowledge which a student would certainly need in order to be successful in the teacher education program with which you are most familiar.
>
> "Important" - of great value to success in a teacher education program. These include skills or knowledge which if present in a student would certainly contribute to a successful experience in teacher education; however it would be possible for a student without some of these skills and knowledge to succeed in teacher education.
>
> "Of Questionable Importance" - these include skills or knowledge which might enhance performance in teacher education but which are not essential to success.
>
> "Irrelevant to Success in Teacher Education" - have no value or effect; these skills or knowledge have no bearing whatsoever on successful performance in teacher education.

Instructional Validity (Opportunity-to-Learn). College and university judges were provided instruction concerning the requirement that students who are required to pass an examination must have had a reasonable opportunity to master the skills or to acquire the knowledge assessed by that examination.

Three types of opportunity-to-learn evidence were obtained: judgments based on the test blueprint, global assessments, and item-focused judgments.

Judges were given a copy of the blueprint for the subtest they were to review. The blueprints for the various subtests contained between three and twelve statements that described the components of the subtests. Judges were asked to answer "yes" or "no" to the following question, for each component of their blueprint:

> "Would 90% or more of the applicants to the teacher education programs at my institution have had a reasonable opportunity to secure instruction in the skills or knowledge defined by each statement on the test blueprint prior to application for admission to teacher education?"

In the global evaluation, judges were asked to describe the degree of similarity between the NTE subtest they reviewed and the relevant part of the general education curriculum at their institution. They used the following four-point scale:

> A. The content of this test and the content included in the relevant portions of my institution's general education curriculum are very similar.
>
> B. There are some differences between the content of this test and the content included in the relevant portions of my institution's general education curriculum, but these differences do not appear to be appreciable.
>
> C. There appear to be some appreciable differences between the content of this test and the content included in relevant portions of my institution's general education curriculum.
>
> D. There appears to be little similarity between the content of this test and the content included in relevant portions of my institution's general education curriculum.

Finally, judges were given a copy of the subtest and the instructional validity response form. They were asked to read each item, to determine the skill or knowledge assessed by the item, and then to answer the following question "yes" or "no:"

> Would 90% or more of the applicants to the teacher education programs at my institution have had a reasonable opportunity to secure instruction in the skills or knowledge represented by this test item prior to applying for admission?

Summary Indices of Relevance and Opportunity To Learn. Relevance and opportunity-to-learn judgments were used to construct four indices of subtest relevance and four indices of examinees' opportunity to learn. Each of those indices is defined in Table 2. Although each relevance index is based on the same data, the indices were construed in different ways. PERRE, the first index, is defined as the percentage of items on the subtest that a judge rated as "crucial" or "important." WMR is defined as the average of a judge's item relevance ratings, using the 4-point scale. BIWMRI is based on a dichotomization of the information in WMR into two categories. SUMRI indicates whether WMR is above a value of 2.25, a commonly-employed threshold for this scale in the item relevance literature. The indices of opportunity to learn (with the exception of BIGLOBAL) are based on different sources of information. GLOBAL is based on judges' global evaluations, and is defined on a four-point scale. INDEXBL is a dichotomously-scored variable that is based on test blueprint judgments. It is scored "1" if a judge answers "yes" to the question described earlier, for at least two-thirds of the categories on a test blueprint. PERIC is based on item-focused judgments; it represents the percent of items for which a judge answered "yes" to the opportunity-to-learn question

described earlier. Summary statistics for the various indices are shown in Table 2.

Insert Table 2 about here

Participants' Evaluation Questionnaire. At the conclusion of the day's validation and standard-setting activities, judges completed a participants' questionnaire designed to secure information on their attitudes toward testing of applicants to teacher education programs, their judgments of the adequacy of the procedures used to elicit their validity and standard-setting judgments, and their personal confidence in the judgments they had provided. The questionnaire contained eleven Likert-style items positively or negatively worded statements with response options on a five-point, "strongly agree" to "strongly disagree" scale. The middle (third) response category was "don't know." Choices in the "don't know" category were considered to be non-responses and were treated as missing values when these data were analyzed.

Since public school judges did not provide judgments concerning opportunity to learn, they were told not to respond to two questions that sought information on these types of judgments. A description of each of the questionnaire items is contained in Table 3.

Insert Table 3 about here

Summary Indices of Judges' Attitudes and Assessments. Prior to data analysis, responses to items on the participants' evaluation questionnaire were aggregated into four variables. Judges' expressed confidence in their

validity ratings were assumed to assess the quality of their ratings. Their assessment of the adequacy of the relevance or instructional validity procedures constituted the second variable. A third variable was composed of responses to items that reflected judges' acceptance of the need to validate the NTE. The fourth variable was composed of items that assessed judges' general attitudes toward screening candidates with standardized tests and the NTE in particular. The definitions of each of these four summary indices are presented in Table 4.

Because distributions of responses to many of the items on the participants' evaluation questionnaire were asymmetrical, responses in the "Strongly Agree" and "Agree" categories were combined, the "Strongly Disagree" responses were combined with the "Disagree" responses, and responses in the "Don't Know" category were treated as missing data.

Insert Table 4 about here

Statistical Analysis. Data were summarized in frequency distributions and measures of central tendency and variability were computed for each variable. Research questions three and four, which asked about relationships between the indices of quality, adequacy, attitudes and acceptance and judgments of relevance and opportunity to learn, were addressed through one-way analyses of variance for continuous validity indices; when validity indices were categorical, chi-square analyses were used. The probability of committing a of Type I error was set at 0.05 for these tests. Measures of association (a correlation ratio, or a contingency coefficient, respectively) were computed to describe strengths of association for the inferential analyses. Log-linear analyses were used to evaluate the

individual and combined effects of predictor variables when sample sizes and variability were sufficient to support such analyses.

Results

Judges' Evaluation of Quality, Adequacy, Attitude, and Acceptance

Frequency distributions of the responses of 244 judges to the eleven items on the Participants' Evaluation Questionnaire are provided in Table 3. The items are grouped into four variables that measure, respectively, the judges' assessment of the quality of their validity recommendations, the judges' perceptions of the adequacy of data collection procedures, the judges' perceptions of the need to validate the NTE, and the judges' attitudes toward testing. Responses to individual items were used to construct indices of "quality," "adequacy," "acceptance of the need for validation," and "attitude toward testing." Those indices were described earlier; distributions of scores on the indices are shown in Table 4. A discussion of these distributions follows.

Confidence in Quality. Almost four-fifths of the judges whose responses could be classified indicated that they were confident about their judgments concerning the relevance of the the NTE subtest they reviewed to success in a teacher education program. About the same proportion indicated that they were confident about their judgments concerning examinees' opportunity to learn the material assessed by the NTE subtest they reviewed.

Adequacy of Procedures. The same pattern of response was found for judges' assessments of the adequacy of the validation procedures that were used. Overall, 77% of the judges responded affirmatively to the items that addressed adequacy of validation procedures. The item that addressed the

difficulty judges had in making relevance judgments elicited the greatest number of negative responses (f=120), however 49% of responses to this question indicated that judges had little or no difficulty. Responses to the other three questions in this set suggest that judges felt the procedures were clear and appropriate. Corresponding aggregate indices reflected the satisfaction of judges with both sets of validation procedures -- 78% and 91% affirmative responses for BADEQREL (relevance procedures) and BADEQIV (instructional validity procedures), respectively.

Acceptance of Need to Validate. The extent to which the judges accepted the need to validate the NTE was assessed by two items on the Participants' Evaluation Questionnaire. Ninety-eight percent of the judges responding (omitting "don't know" and missing responses) agreed that there was such a need. BACCEPT, the aggregate index also had a 97% affirmative response. Lack of variability in the BACCEPT index was problematic, and limited subsequent analyses of the relationship between this variable and corresponding validity judgments.

Attitude Toward Testing. Judges' reported attitudes toward testing varied more than did their responses to most of the other items on the Participants' Evaluation Questionnaire. Overall, about 79% of the respondents to the three items measuring such attitudes agreed that it was important to screen applicants to teacher education programs and that standardized tests, including the NTE, were appropriate mechanisms to use in such screening; however 21% of respondents disagreed with those sentiments. Results for the aggregate index (BATTIT) were consistent with the distributions of responses to items that compose it.

Judges' Validity Recommendations

Judges' recommendations regarding the relevance of the skills measured by NTE to successful performance as a teacher education student and their judgments regarding examinees' opportunity to learn were used to compute four aggregate indices of relevance and four indices of examinees' opportunity to learn. These indices are defined in Table 2. The means and standard deviations of continuously-distributed indices and the frequency distributions of dichotomous indices are also shown in Table 2.

Both college and university judges and public school judges were asked to provide judgments that contributed to computation of the relevance indices. However, only college and university judges were asked to provide judgments that contributed to computation of the opportunity-to-learn (instructional validity) indices.

In general, the assessments of validity were positive. For example, 72% of the responses characterized the knowledge and skills assessed by the NTE subtests as either "crucial" or "important" to success in teacher education. On average, 79% of items concerned with examinees' opportunity to learn were answered affirmatively. The statistics shown in Table 2 suggest that some of the indices were more variable than others. For example, lack of variance in the SUMRI (relevance) and INDEXBL (opportunity to learn) limited analyses of relationships between these indices and judges' evaluations of the adequacy of validation procedures and their confidence in the judgments they provided.

Relationships Between Types and Perceptions of Judges and Validity Ratings

Research Questions 3 and 4 asked whether there were relationships between various judge-related variables (defined in Table 4) and judges'

assessments of the relevance of NTE subtests and examinees' opportunity to learn. Analyses of these relationships are summarized in Table 5.

Insert Table 5 about here

Relationships between the variables type of judge, confidence (quality), adequacy, attitude and acceptance, and each of the relevance and instructional validity indices were examined using either analyses of variance or Chi-square analyses. Judges' evaluations of their confidence and of the adequacy of validation procedures were secured after they had assessed the relevance of the NTE subtests and examinees' opportunity to learn.

Twenty-seven statistical tests were conducted, of which 12 were found to be significant, using a Type I error level of 0.05. At this Type I error level, fewer than two tests would be expected to result in rejection of the null hypothesis just due to sampling error. Since more than 40 percent of the tests produced significant results and there appears to be a consistent pattern of results for the dependent variables investigated, it is highly unlikely that the observed results can be attributed to the effects of a large family error rate (Miller, 1966).

Type of Judge. No overall differences in validity assessments were found when judgments provided by the two types of judges were compared.

Confidence in Quality. Judges' confidence in their relevance assessments were significantly related to each of the three indices of the relevance of the NTE subtests. Judges who were more confident were more likely to rate the NTE items as relevant to successful performance in a teacher education program. Measures of association for the significant

effects were 0.18, (a contingency coefficient), 0.22 and 0.26 (correlation ratios). Judges' expressions of confidence in their opportunity to learn recommendations were not related to their opportunity-to-learn validity assessments.

Adequacy of Procedures. Judges' assessments of the adequacy of validation procedures were not found to be related to any of their assessments of the relevance of the NTE, nor were their assessments of the adequacy of validation procedures significantly related to any of their assessments of examinees' opportunity to learn.

Judges' Attitudes. Attitudes of judges were found to be significantly related to their evaluations of the validity of the NTE; five of six tests of association produced statistically significant results. However, indices of strength of relationship were generally of modest size. Correlation ratio values associated with significant F-statistics were 0.22, 0.32, 0.33, and contingency coefficients associated with significant Chi-square analyses were 0.20 and 0.36. Judges who believed that it was important to screen candidates with standardized tests, including the NTE, were more likely to assess the NTE subtests as relevant, were more likely to judge that the content of subtests and curricula were similar, and were more likely to judge that examinees had an opportunity to learn the skills specified in the subtest blueprints.

Judges' Acceptance of the Need to Validate. Judges' acceptance of the need to validate the NTE was significantly related to four of the six validity indices. The strength of relationship was modest in each instance. Correlation ratios associated with two significance tests were 0.20 and 0.21 contingency coefficients associated with two others were 0.23 and 0.26. Judges who agreed with the need to validate the NTE were more likely to

judge the test items as relevant, and were more likely to have judged that examinees would have had an opportunity to learn the skills represented in the NTE items and test blueprints. They were also more likely to have agreed that the content of an NTE subtest and the content of the teacher education curriculum at their institution were similar. These results should be regarded with some caution because the numbers of judges in the "doesn't accept" cells were quite small.

A Possible Rival Hypothesis: Was There Confounding Due to Subtest Differences?

Data obtained from judges who evaluated different NTE subtests have been pooled in the analyses described to this point. If the assessments of judges who evaluated different NTE subtests had differing relationships to the judges' perceptions of procedural adequacy, etc., the pooled results would be subject to confounding effects and might be susceptible to misinterpretation. For these reasons, a preliminary set of comparisons was performed to determine whether there were between-subtest differences in judges' NTE validity assessments or their evaluations of procedural adequacy, the quality of their judgments, etc. Relationships between NTE subtest, the validity indices and judges' evaluation indices were tested using either one-factor analyses of variance or Chi-square analyses. The results of these analyses are presented in Table 6.

Insert Table 6 about here

Subtest Differences in Validity Judgments. Significant between-subtest differences were found for five of the six relevance and opportunity-

to-learn indices. Table 6 contains descriptive statistics for each subtest. The Reading and Writing subtests received the highest relevance ratings. It appears that the content assessed by the Science subtest is regarded as being less relevant to success in teacher education programs than is the content assessed by the other tests. There is considerable variation among the NTE subtests with regard to opportunity-to-learn ratings. These indices are based on different sources of data (global, blueprint and individual items) and would be expected to be more independent of each other than would the relevance indices.

Subtest Differences in Judges' Evaluations. Four of the six indices of judges' evaluations of quality, adequacy, attitudes and acceptance were found to differ significantly among subtests (see Table 6). Judges' who evaluated the Listening subtest reported the lowest level of confidence in their recommendations and the most negative assessment of the adequacy of validation procedures. However all indices vary considerably among the various subtests.

Based on these preliminary analyses we concluded that confounding of relational results due to subtest differences was a possibility. We therefore attempted to assess and control for the effects of this source of variance. Two sets of analyses were completed. First, separate analyses were completed using the data provided by judges who evaluated each NTE subtest. This analytic strategy was problematic because sample sizes were reduced, with consequent losses in statistical power. Furthermore, if statistically significant relationships between variables were found for some but not all of the subtests, the generalizability of results would be restricted. The second approach was to conduct a series of log-linear analyses that incorporated a subtest by dependent variable interaction term in their

design. This approach had the added advantage of providing an investigation of interactions among the variables. However, it had the disadvantage of requiring a sufficient number of observations in all cells of the contingency tables used in the analyses.

Separate Analyses: Control for Subtest Differences

Analyses of relationships among type of judge, judges' assessments of the validity of the NTE subtests, and judges' evaluations of the procedures used to elicit validity judgments, etc. were conducted separately, using data secured for each of the NTE subtests. The analytic procedures used were identical to those described earlier, with the following exception: For small samples, Fishers Exact Test was employed as an alternative to the Chi-square test of association. If subtest differences were responsible for the total-group associations reported earlier, these associations would be expected to disappear in the current analyses. The results of analyses based on data for individual NTE subtests are summarized in Table 7.

In order to keep the statistical power for these hypothesis tests comparable to that of the analyses described earlier, the probability of committing a Type I error was raised to $p < .10$. Unfortunately, stratification of the data by NTE subtest resulted in very small samples, and in some cases made statistical tests impossible.

Results shown in Table 7 summarize the number of statistical tests that resulted in rejection of the null hypothesis of no association between each validity index and each "quality" index, for the set of seven multiple-choice NTE subtests. The numbers of instances in which there were sufficient cases to perform a significance test are also shown in Table 7. In interpreting these results, the relatively large family error rate associated with the entire

set of statistical tests should be kept in mind. One would expect that as many as 10% of the significant results might be attributable to Type I errors. For this reason, we have summarized the results of these tests by showing the number of significant outcomes and the actual number of statistical tests performed. When the proportion of statistically significant outcomes approached 0.10, the results were discounted as probably attributable to Type I error. When the proportion of significant outcomes exceeded 0.10, we searched for, and noted, consistent patterns in the results.

Insert Table 7 about here

In the analysis of pooled data reported earlier, type of judge was not found to be systematically related to validity recommendations. Identical results were found when data were analyzed separately by NTE subtest. In the earlier analysis of pooled data, judges' confidence in their relevance judgments was found to be related to their assessments of the relevance of NTE subtests to success in teacher education programs. When data were analyzed separately by subtest, these findings were replicated for only two of the seven NTE subtests: the Writing subtest and the Listening subtest. In the earlier analysis of pooled data, no relationships were found between judges' assessments of the adequacy of the relevance judgment procedures and their judgments of the relevance of the NTE subtests. Similar results were obtained when data were analyzed separately by NTE subtest.

Neither sets of judgments (confidence and adequacy) regarding opportunity-to-learn activities were found to be related to opportunity-to-learn assessments of the NTE in the earlier pooled analyses. When data were

analyzed by separate NTE subtest, no consistent pattern of relationships was found.

In the earlier pooled analyses judges' attitudes were found to be systematically related to their relevance and opportunity assessments. When data were analyzed by NTE subtest, judges' attitudes were found to be significantly related to their validity assessments in 11 of 33 contrasts. These results support the conclusion that judges' attitudes are related to their relevance and opportunity-to-learn judgments for the Reading, and Science subtests.

Examination of relationships involving judges' acceptance of the need to validate screening tests was impeded by lack of variance. However, based on very limited data, we conclude that judgments of the relevance of the Science subtest might be related to judges' acceptance of the need to validate the NTE. These results must be considered cautiously, since available data were limited and were asymetrically distributed.

Results of the Loglinear Analysis

Logistic modeling procedures were used to predict two dichotomous dependent variables selected to measure assessed relevance of the NTE to successful performance in teacher training (BIWMRI) and to measure opportunity-to-learn the content of the NTE subtest (BIGLOBAL). Each dependent variable was predicted by two sets of categorical independent variables. The prediction equations for each dependent variable contained the following possible predictors:

Equation	Dependent Variable	Possible Predictors
1	BIWMRI	type of judge, confidence, adequacy, subtest
2	BIWMRI	type of judge, attitude, acceptance, subtest
3	BIGLOBAL	confidence, adequacy, subtest
4	BIGLOBAL	attitude, acceptance, subtest

These analyses replicated the testing of main effects in the analyses described earlier, but included the possibility of interaction effects. The analyses also allowed us to examine subtest interaction effects and thereby address the question of confounding in the pooled results. In each of the models tested all interactions with subtest were removed following examination of a model that included them, because they did not contribute significantly to explained variance. This finding eliminates the possibility of confounding by subtest, at least for the dependent variables examined here. Results of the logit analyses are described for each dependent variable.

Relevance. The model which best explained observed variation consisted of the following effects: Main effects for confidence and subtest. The Chi square fit statistic was 30.02 (df=29) an associated fit probability of .413. Estimated coefficients were used to calculate the expected odds of judging each of the subtests as "relevant". The results of that analysis are shown in Figure 1.

Insert Figure 1 about here

Not only were different odds found for the various NTE subtests, but judges' confidence in their ratings was found to be substantially related to odds. Overall, there is significant likelihood that the subtests will be judged relevant, as opposed to irrelevant. The Science and Social Studies subtests have the smallest expected odds of being judged relevant. The expected odds of a subtest being judged relevant increased two- or three-fold when the judges were confident about their rating.

Opportunity-To-Learn. The dependent variable (BIGLOBAL) measured the judged degree of global similarity between the content of the teacher education curricula and an NTE subtest. Since only college and university judges provided such judgments, type of judge was not included in the analysis of this variable. The model which best explained observed variation in BIGLOBAL consisted of the following effects: The main effects for judges' confidence and subtest. The Chi square fit statistic was 16.01 (df=12) with an associated fit probability of .191. Estimated coefficients were used to calculate the expected odds of judging each of the subtests as "similar" when judges' confidence ratings were considered. The results of that analysis are shown in Figure 2. Overall, more judges rated the NTE as "similar" than "dissimilar", but a null hypothesis of equal population proportions would be retained. There are differences in expected odds when the different

Insert Figure 2 about here

subtests were considered. For the Writing subtest a higher expected odds of eliciting a "similar" judgment were found. The Social Studies subtest was associated with the smallest expected odds of "similar" judgment. In this analysis, whether the judge was confident or not had a very large association with expected odds. In the case of the Writing subtest, the odds of a "similar" response increased from about 4:1 to 9.5:1 when the responding judge was confident. The effect was present across all NTE subtests.

The model which best explained observed variation in similarity judgments as a function of attitude variables consisted of the following effects: The main effects for judges' attitude and subtest. The Chi square fit statistic was 10.01 (df=7) with an associated fit probability of .183. The expected odds of judging each of the subtests as "similar" when the judges' confidence ratings were considered, are shown in Figure 3.

Insert Figure 3 about here

In this analysis, the expected odds of observing a "similar" response for the Writing subtest were greater than found for the other tests. The attitudes of judges' toward screening and toward the use of standardized tests were highly associated with the odds of a subtest being rated "similar". Judges who believed in testing and use of the NTE were found to be more likely to give a positive response to opportunity to learn questions. For the Writing subtest the expected odds increased from 2 :1 to 7.6 : 1, for the judges who had a positive attitude.

Discussion

Not all judges agree that standardized tests, and the NTE in particular, are the best way to screen teacher education applicants, in fact, some judges do not agree that applicants should be screened at all. It is interesting to note that, despite judges' positions regarding selection procedures, most agree that if such tests are used, they should be validated before they are used.

Judges' assessments of the relevance and examinees' opportunity to learn the content of the NTE were generally positive, but there were differences between judges' evaluations of individual subtests. The Science subtest was rated lowest with respect to relevance to success in teacher education programs. This finding suggests that this subtest, as well as the others, should continue to undergo periodic review.

Overall, judges serving on the panels in this study were reasonably confident about the quality of their relevance and opportunity-to-learn recommendations and they judged the data collection procedures used in this study to be adequate. Procedures used to assess examinees' opportunity to learn were evaluated somewhat more positively than were procedures used to assess the relevance of the NTE. There were significant differences among evaluations of the procedures by judges who assessed different NTE subtests; in particular, the lowest confidence ratings were provided by judges who assessed the Listening subtest.

The answer to the question of whether relevance and opportunity-to-learn ratings are related to judges' characteristics and self assessments has not been answered clearly. Correlations were not found to be consistent for all subtests. Although many of the correlations diminished when the data were analyzed by subtest, some findings suggest that such relationships exist for certain of the subtests.

Judges who rate subtests as relevant are more likely to be confident about their recommendations, and are more likely to have a positive attitude toward the use of standardized tests and the NTE. The findings from the logist analyses also suggest that, for the two dependent variables studied, there may be systematic relationships between validity ratings and certain characteristics of judges, such as their attitudes and their confidence in their ratings. This finding is potentially troubling. One must wonder about the implications of those observations for the validity of content validity evidence judges provide since these sources of systematic variation are properties of judges rather than the subtests that are the purported objects of judgment.

Future studies of these issues might be improved by incorporating several methodological modifications. The evaluation questionnaire could be improved in two ways. Several variables were assessed by a single item. Additional items should be written to assess these variables and thereby increase measurement reliability. Certain of the items, notably those measuring acceptance of the need to validate the NTE produced little variance. Perhaps alternative items that more sensitively reflect variability in these attitudes can be found. In addition, future studies should incorporate larger samples of judges and thereby support statistical tests with greater power.

Judges are central to procedures designed to assess the relevance of tests and examinees' opportunity to learn the content of tests. Although the results reported here are somewhat equivocal, we have found disturbing relationships between variables that reflect characteristics of judges and variables that reflect judges' assessments of the quality of tests. These findings underscore the need to carefully and continually examine the

procedures that enhance judges' confidence in the quality of their judgments and their perceptions of procedural adequacy.

References

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Brown, E. & Rankin, R. (1986). Predicting employment in education: The relative efficiency of National Teacher Examinations scores and teacher ratings. Educational and Psychological Measurement, 46, 191-197.

Busch, J. C. (1988). Factors influencing recommended standards for essay tests. Applied Measurement in Education, 1(1), 67-78.

Busch, J. C. & Jaeger, R. M. (1986). Judges' background, attitudes and information as concomitants of recommended test standards. Paper presented at the Annual Meeting of the American Educational Research Association. Resources in Education (ED271 482).

Busch J. C. & Jaeger, R. M. (1988). Influence of type of judge, normative information and discussion on standards recommended for the National Teacher Examinations. Unpublished manuscript.

Carlson, R. E. (1988). Alternative procedures for collecting and analyzing job relatedness judgments in NTE test appraisal studies. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Cross, L. (1985). Validation of the NTE tests for certification decisions. Educational Measurement: Issues and Practice, 4, 7-10

Cross, L. (1988). Reservations and recommendations regarding the validation of the NTE program tests. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Cross, L., Frary, R. B., Kelley, P.P., Small, R. C., Impara, J. C. (1985). Establishing minimum standards for essays: Blind vs. informed reviews Journal of Educational Measurement, 22 (2), 137-146.

Equal Employment Opportunity Commission. Civil Rights Commission. Department of Labor & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43 (166), 38290-38315.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-475.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 485-514). New York: American Council on Education and Macmillan.

Jaeger, R. M. Busch, J. C. (1984). A validation and standard-setting study of the General Knowledge and Communication Skills tests of the National Teacher Examinations, Center for Educational Research and Evaluation: University of North Carolina at Greensboro.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Miller, R. G. (1966). Simultaneous statistical inference. New York: McGraw-Hill.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 335-366). New York: American Council on Education and Macmillan.

Poggio, J., Miller, M., & Glasnap, D. (1987). The adequacy of retrospective judgments to establish instructional validity. Educational and Psychological Measurement , 47, 783-792.

Rosenfield, M., Thornton, R. F., & Skurnik, L.S. (1986). Analysis of the professional functions of teachers. Relationships between job functions and the NTE Core Battery. Princeton: Educational Testing Service. Resources in Education (ED275 693).

Rosner, F. & Howey, K. (1982). Construct validity in assessing teacher Knowledge: New NTE interpretations. Journal of Teacher Education, 33 (6), 7-12.

Rudner, L. (August, 1987). What's happening in teacher testing. U.S. Department of Education. Washington: Superintendent of Documents, 145 pp.

Table 1.
Number of Judges on Each Panel Who Were Public School or College/University Faculty.

Type of Judge	NTE Communication Skills Subtests		
	Reading	Writing	Listening
College/University	21	36	25
Public School	5	14	18
Total	26	50	43

Type of Judge	NTE General Knowledge Subtests			
	Mathematics	Science	Social Studies	Literature/ Fine Arts
College/University	17	23	23	24
Public School	8	9	8	14
Total	25	32	31	38

Table 2
Definition and Descriptive Summary of Relevance and Opportunity-to-Learn Indices.

Index Name	Definition of Index	Type of Variable	Summary Statistics		
Relevance Judgments			X	s=	n=
PERRE	Percent of crucial/important judgments on the test.	Contin	72.46	20.49	242
			X	s=	n=
WMR	Mean on 4-pt rating scale (crucial to not relev.)	Contin	2.98	.50	242
			0=not rel.		1=rel.
BIWMRI	Dichot. categorization of WMR variable.	Categ	f=36		f=206
SUMRI	Whether judges aver. relev. rating was greater than 2.25 threshold or not.	Categ	0=LT Thresh f=18		1=GE Thresh f=224
Opportunity-to-Learn Judgments (Instructional Preparation)			X	s=	n=
PERIC	Percent items on test answered "yes"	Contin	79.38	22.97	162
INDEXBL	2/3 or more of blue-print categories answered "yes"	Categ	0=LT 2/3 "yes" f=25		1=GE 2/3 "yes" f=132
Opportunity-To-Learn (Curricular Coverage)					
GLOBAL	Similarity of Test	Categ	Little Sim. f=7	Some Diff f=56	No Ap. f=72	Very Sim. f=27
BIGLOBAL	Dichotomous version of Global Variable	Categ	Not Similar f=63			Similar f=99

Table 3.
Frequency Distributions of 244 Judges' Responses to Evaluation Items.

	RESPONSE CATEGORY					
Judges Confidence in Quality of Recommendation	**SD**	**D**	**DK**	**A**	**SA**	**NR**
10. Feel confident about Relev. judgments.	6	44	28	131	30	5
25. Confident about Instr. Val.. ratings.	5	24	20	81	34	80
Judges' Assessment of Adeq. of Procedures						
8. Didn't fully understand Relev. directions. (R)	72	124	3	34	7	4
9. Difficult to make Relev. judgments. (R)	18	97	5	96	24	4
11. Had enough time for Relev. judgments.	5	20	5	138	72	4
24. Instr. Valid. procedures explained clearly.	2	13	6	92	52	79
Judges Value Need to Validate						
1. Understand need to validate NTE	0	3	0	82	155	4
3. Important to determine Relev. of NTE.	4	3	2	76	154	5
Judges Attitude Toward Testing						
4. Not important to screen candidates. (R)	159	57	9	7	8	4
5. Standard. testing is appropriate to screen candidates.	19	36	42	114	29	4
6. NTE tests are not appropriate to screen.(R)	37	98	39	40	25	5

(R) denotes a negatively stated item for which the scale was reversed.

Table 4.

Summary of Indices of Judges' Evaluation of Confidence, Adequacy Attitudes and Acceptance.

Confidence in Quality of Recommendations

Index: BCONFREL*
Definition: Confidence in relevance judgments.
Computed: Response to item 10 — Confident(1) n=161; Not Conf. (0) n= 50

Index: BCONFIV*
Definition: Confidence in opportunity-to-learn ratings.
Computed: Response to item 25 — Confident (1) n=115; Not Conf. (0) n= 29

Adequacy of Procedures

Index: BADEQREL*
Definition: Adequacy of procedures for relevance judgments.
Computed: Response to items: 8, 9, 11 — Adequate (1) n=179; Inadeq (0) n= 49

Index: BADEQIV*
Definition: Adequacy of procedures for opportunity-to-learn ratings.
Computed: Response to item 24 — Adequate (1) n=144; Inadequate (0) n= 15

Attitude Toward Testing

Index: BATT
Definition: Appropriateness of standardized tests (NTE) to screen candidates.
Computed: Response to items: 4, 5, 6 — Positive (1) n= 138; Negative (0) n= 35

Accepts Need and Importance of Validation of the NTE

Index: BACCEPT
Definition: Agrees that screening tests should be validated.
Computed: Response to items:1 and 3. — Accepts(1) n=230; Not Accept(0) n= 7

* The sample sizes for the opportunity-to-learn indices are less than the relevance indices because they were based solely on recommendations from college/university judges. As a consequence of the fact that some items were not answered by all judges, indices which were composites of several items were "missing". The "don't know" (3) category was scored as missing.

Table 5.
Summary of Evidence Regarding The Relationship Between Type of Judge, Quality (Confidence), Adequacy, Attitude, Acceptance Indicators With Validity Indices.

Independent Variables:	Validity Indices: Relevance			Instructional Validity		
	PERRE	WMR	SUMRI	PERIC	INDEXBL	GLOBAL
Type of Judge	NS	NS	NS			
Confidence in:						
(BCONFREL)	**	**	*			
Instr. Valid J. (BCONFIV)				NS	NS	*
Adequacy Proc. Relev Judge. (BADEQREL)	NS	NS	NS			
Instr. Valid J. (BADEQIV)				NS	NS	NS
Attitude:	**	**	*	NS	**	*
Acceptance:	NS	NS	**	**	**	*

* $p < .05$
** $p < .01$
NS : Not Significant
Blanks: Contrast not made because it does not make theoretical sense.

Table 6.

Comparisons Across Subtests for Validity Indices and Judges Evaluations.

Validity Indices

Index	Significant Difference	Mean for NTE Subtest: Rea	Wri	Lis	Mat	Sci	Soc	Lit
PERRE	**	84.0	77.0	74.8	70.9	59.0	69.1	71.1
WMR	*	3.29	3.07	3.04	2.96	2.62	2.93	2.97
PERIC	*	86.6	85.9	82.5	85.8	68.2	73.0	72.8
GLOBAL	**	1.95	1.58	2.54	2.47	2.52	2.59	2.52
		% of Index Values = "1" for NTE Subtest a.						
		Rea	Wri	Lis	Mat	Sci	Soc	Lit
SUMRI	NS	100%	96%	90%	92%	81%	93%	95%
INDEXBL	*	86%	100%	79%	88%	91%	62%	77%

Evaluation Indices

Index	Significant Difference	% of Index Values = "1" for NTE Subtest: a. Rea	Wri	Lis	Mat	Sci	Soc	Lit
BCONFREL	**	96%	72%	51%	81%	77%	81%	93%
BCONFIV	*	79%	76%	57%	93%	88%	78%	90%
BADEQREL	**	85%	65%	59%	96%	90%	92%	91%
BADEQIV	NS	100%	94%	73%	88%	95%	91%	90%
BATTIT	*	72%	81%	97%	76%	63%	67%	88%
BACCEPT	NS	100%	98%	98%	100%	90%	100%	94%

* p< .05

**p< .01

NS= not significant

NTE Subtests: Reading (Rea), Writing (Wri), Listening (Lis), Mathematics(Mat), Science (Sci), Social Studies (Soc), Literature/Fine Arts (Lit).

a. These are dichotomous variables; the definitions of a "1" value are given in Tables 2 and 4.

Table 7.
Summary of Significance Tests Performed on Individual Subtests.

Independent Variables:		Validity Indices: Relevance PERRE	WMR	SUMRI	Instructional Validity PERIC	INDEX	BL GLOBAL
Type of Judge	NTP:	7	7	7			
	NSR:	0	0	0			
	SSR:						
Confidence in:							
Relev. Judge	NTP:	7	7	6			
(BCONFREL)	NSR:	2	2	0			
	SSR:	2,3	2,3				
Instr. Valid J.	NTP:				6	7	6
(BCONFIV)	NSR:				1	0	2
	SSR:				7		1,2
Adequacy Proc.							
Relev Judge.	NTP:	7	7	6			
(BADEQREL)	NSR:	7	1	0			
	SSR:	3	3				
Instr. Valid J.	NTP:				5	6	5
(BADEQIV)	NSR:				1	0	0
	SSR:					6	
Attitude:	NTP:	6	5	5	6	5	6
	NSR:	3	3	1	0	2	2
	SSR:	1,4,5	1,4,5	5		1,6	1,2
Acceptance:	NTP:	3	3	4	2	3	1
	NSR:	1	1	1	1	0	0
	SSR:	5	5	5	5		

Blanks: Contrasts not made because they do not make theoretical sense.
NTP: Number of actual tests performed
NSR: Number of significant results
SSR: Subtests with significant results: (1) Reading, (2) Writing, (3) Listening, (4) Mathematics , (5) Science, (6) Social Studies, (7) Literature/Fine Arts.

Figure 1. Expected odds of "relevant" ratings by confident and non-confident judges.

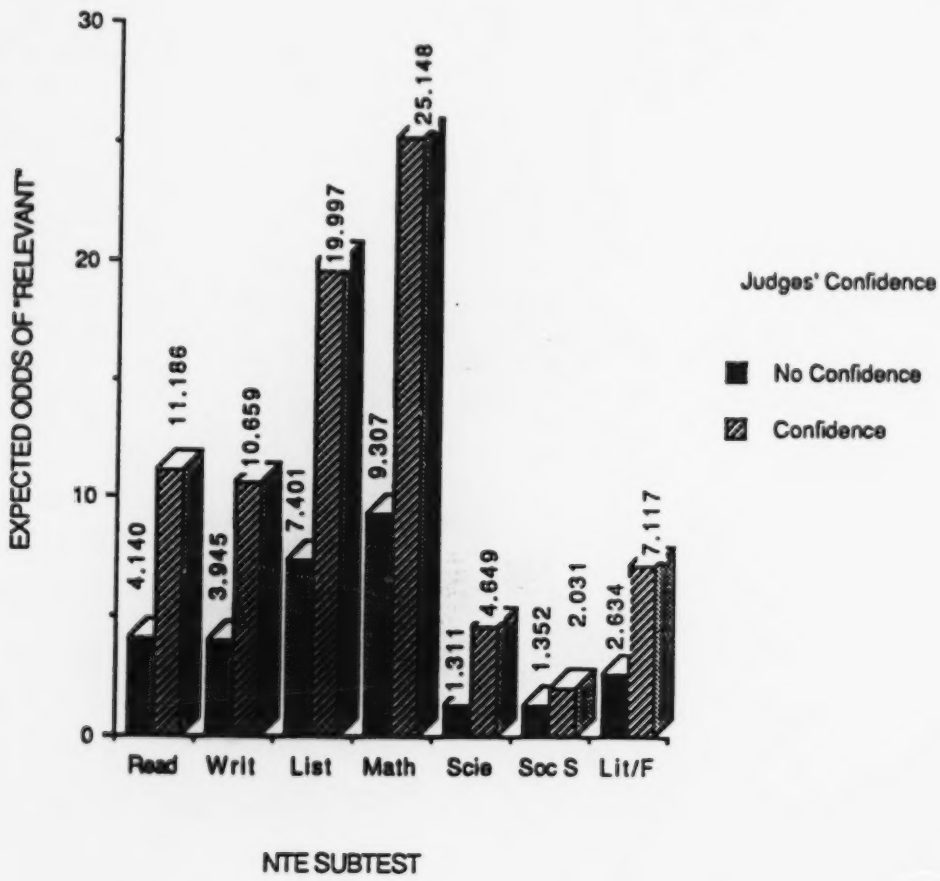

Figure 2. Expected odds of "similar" ratings by confident and non-confident judges.

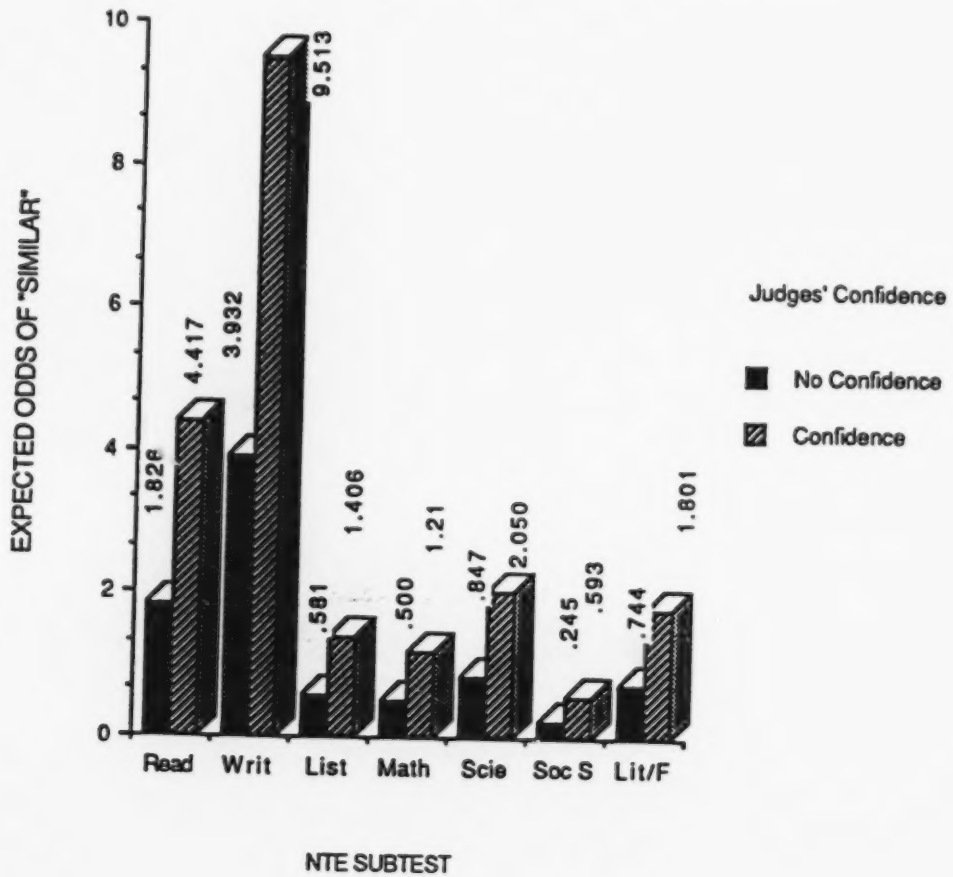

Figure 3. Expected odds of "similar" ratings by judges with positive and negative attitudes.